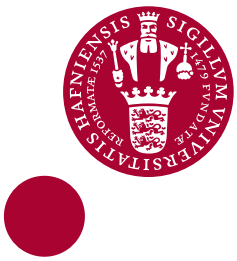


Texture side information generation for distributed video coding of video-plus-depth

Salmistraro, Matteo; Rakêt, Lars Lau; Zamarin, Marco; Ukhanova, Ann; Forchhammer, Søren

Published in:
2013 20th IEEE International Conference on Image Processing (ICIP)

DOI:
10.1109/ICIP.2013.6738350

Publication date:
2013

Document version
Early version, also known as pre-print

TEXTURE SIDE INFORMATION GENERATION FOR DISTRIBUTED CODING OF VIDEO-PLUS-DEPTH

Matteo Salmistraro[◊] *Lars Lau Rakêt*[⋆] *Marco Zamarin*[◊] *Anna Ukhanova*[◊] *Søren Forchhammer*[◊]

[◊]DTU Fotonik, Technical University of Denmark, Ørsteds Plads,
2800 Kgs. Lyngby, Denmark. Emails: {matsl, mzam, annuk, sofo}@fotonik.dtu.dk
[⋆]Department of Computer Science, University of Copenhagen, Universitetsparken 5,
2100 Copenhagen, Denmark. Email: larslau@diku.dk

ABSTRACT

We consider distributed video coding in a monoview video-plus-depth scenario, aiming at coding textures jointly with their corresponding depth stream. Distributed Video Coding (DVC) is a video coding paradigm in which the complexity is shifted from the encoder to the decoder. The Side Information (SI) generation is an important element of the decoder, since the SI is the estimation of the to-be-decoded frame. Depth maps enable the calculation of the distance of an object from the camera. The motion between depth frames and their corresponding texture frames (luminance and chrominance components) is strongly correlated, so the additional depth information may be used to generate more accurate SI for the texture stream, increasing the efficiency of the system. In this paper we propose various methods for accurate texture SI generation, comparing them with other state-of-the-art solutions. The proposed system achieves gains on the reference decoder up to 1.49 dB.

Index Terms— Distributed Video Coding, Depth Map, Wyner-Ziv Coding, Optical Flow, Multi-Hypothesis

1. INTRODUCTION

In the recent years Distributed Video Coding (DVC) has received a great amount of interest, due to the possibility of shifting complexity from the encoder to the decoder.

In this paper we address DVC of video-plus-depth streams in a monoview scenario and propose methods to exploit the correlation between the streams in order to produce more accurate Side Information (SI). Depth maps can be used in single view scenarios for activity detection, object tracking and background/foreground separation [1].

DVC is based on two information theoretic results, the Slepian-Wolf [2] and Wyner-Ziv [3] (WZ) theorems, where, in the second case, source data are independently lossy coded but jointly decoded using a correlated source at the decoder, commonly referred to as SI. DVC could be an appealing solution for the video-plus-depth coding problem, in particular if we require low-complexity encoders. It is possible, in this way, to independently code the two streams and then jointly decode them. This is especially convenient when separated texture and depth cameras are used, in which case inter-camera communication is difficult or perhaps infeasible. The DVC decoder used as basis of our system is the one presented in [4], employing the approach first proposed in [5] and then improved in [6]. As can be seen in Fig. 1, the frames are divided into Key-Frames (KFs) and WZ frames at the encoder. The KFs are encoded independently with respect to each other and with respect to the WZ frames, using a H.264/AVC Intra coder. The KFs are used at the decoder to calculate the SI, which is a prediction of the to-be-decoded WZ frame. At the encoder the WZ frame is DCT-transformed, the coefficients are grouped and divided in bitplanes. Each bitplane is encoded using an LDPCA encoder [7], and a subset of the calculated syndromes is sent to the decoder. The decoder uses the syndromes to correct the errors in the corresponding SI bitplanes, bitplane by bitplane. If the syndromes are not enough, others are requested via a feedback channel. The LDCPA decoder also requires the calculation of the reliability of the bits of the bitplanes. Ideally, it is possible to calculate such reliability from the residual, which is the difference between the SI and the original WZ frame, but since WZ frames are not available at the decoder, a residual estimation method have to be devised.

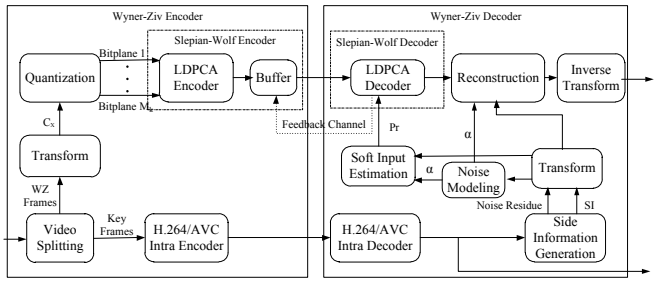
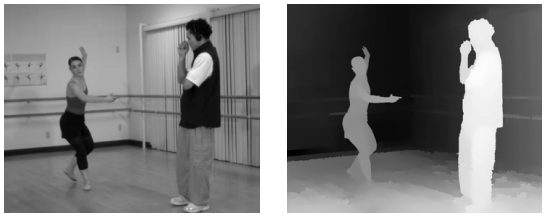


Fig. 1: DVC Codec [4].

Depth maps are images allowing the calculation of the distance of an object from the camera. While texture frames contain the luminance and chrominance components of the



(a) Texture Frame (b) Depth Frame

Fig. 2: A texture frame (a) and its correspoding depth frame (b), from the Ballet [12] sequence.

scene (Fig. 2a), depth maps describe depth information (Fig. 2b). Depth information can be used to calculate the distance of a given point in the 3D scene from the camera. The depth and texture frames of the same scene, referring to the same time instant, are strongly correlated and the motion between texture frames is highly correlated with the motion between the depth frames [8]. This gives rise to the video-plus-depth coding problem, in which the redundancy between the two streams is used to achieve efficient coding, as for example in [8]. This approach has also been used in depth map coding architectures based on DVC [9, 10] where the depth motion estimation has been carried out exploiting texture data. In DVC for texture frames, depth data have been used to generate intra-view SI through view synthesis [11]. View synthesis can be used in Multiview DVC but it is not suitable for single view systems or in the cases in which the Rate-Distortion (RD) performance of the intra and inter-view SIs are too different to obtain improvements from the fusion of the SIs.

This paper proposes methods for exploiting depth maps in the texture SI generation. We consider that an independently coded depth stream is already available and it is used to improve the WZ coding performance of the texture stream. We introduce Optical Flow (OF) based techniques, extending the framework proposed in [13] and introducing a new OF technique based on two distinct data terms. We benchmark these techniques against well-known block-based systems. Finally we consider the use of a multi-hypothesis decoder [14] for efficient and robust SI fusion. OF-based SI generation [15, 16] has been previously used in DVC as a way to create accurate SI for texture streams. In this paper we use OF to extract accurate motion estimates from the depth stream. We also propose a joint stream calculation, taking into account, at the same time, both KFs and depth frames, employing an OF formulation with two constraints.

2. SIDE INFORMATION GENERATION

Let D_i and T_i, with temporal index i, denote depth and texture frames, respectively. The to-be-decoded frame is T_t, all the other frames in Fig. 3 are assumed to be known at the decoder. We estimate the motion between D_t and D_{t-1} and use it to motion compensate T_{t-1} obtaining Y_{t-1}. We also calculate the motion between D_t and D_{t+1}, then T_{t+1} is motion compensated obtaining Y_{t+1}.

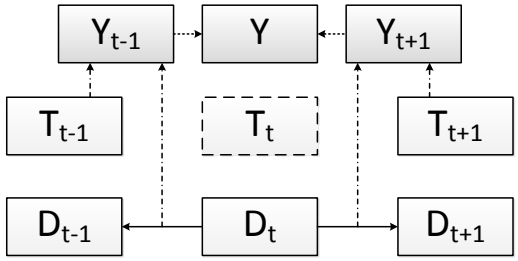


Fig. 3: The video stream structure, Group-Of-Pictures 2.

Once these two components have been calculated, the final SI Y can be calculated as their average, and the residual R can be calculated as their difference. We propose three new methods for this basic setup (Fig. 3).

The first two, "D2T BB" and "D2T OF", calculate the motion using the depth frames only, then this motion is used to motion compensate the texture frames, generating Y_{t-1} and Y_{t+1}. The difference between the two is the Motion Estimation (ME) algorithm: D2T BB uses a Block-Based (BB) method, while D2T OF uses an OF method.

For what concerns D2T BB, we consider the so-called "Adaptive Rood Pattern Search" (ARPS) ME algorithm [17]. While this approach may not provide the lowest Mean-Squared-Error between the motion compensated depth frame and the original one on average, it is able to capture the motion between the frames in a robust way, leading to fewer artefacts in the warped (texture) frame. ARPS has been proposed as a way to reduce the complexity of the ME process in state-of-the-art predictive coding, but thanks to the adaptive nature of the pattern and the refinement step, it produces superior results compared with full search ME in the given setup. The final method that we propose, "DT2T", is an OF method that uses both texture and depth information. This method employs the symmetric (texture) data term proposed in [13], but also adds the asymmetric information given by the depth maps in the motion estimation. In addition, we consider the symmetric texture based SI generation method presented in [16], which produces state-of-the-art results, as an alternative that does not use depth maps. This method is denoted as "T2T".

2.1. Optical Flow based SI generation

As opposed to BB motion estimation, OF gives a dense result, calculated by means of a global regularization process. Typical SI generation methods are based on calculating motion using texture KFs [14, 16]. Here we extend the symmetric OF method of [13] to also include asymmetric depth information. A novelty of our approach is the introduction of a new OF-based SI generation system, in which two data terms are jointly minimized.

Given a set of pixel-domain (texture) key frames and depth frames T_{t-1}, T_{t+1}, D_t, and $D_{t'}$, $t' = t-1$ or $t' = t+1$,

we want to estimate the dense flow field v such that the following optical flow constraints

$$C_T(\boldsymbol{x}, v) \triangleq T_{t+1}(\boldsymbol{x} + v(\boldsymbol{x})) - T_{t-1}(\boldsymbol{x} - v(\boldsymbol{x})), \quad (1)$$

$$C_D(\boldsymbol{x}, v) \triangleq D_{t'}(\boldsymbol{x} + v(\boldsymbol{x})) - D_t(\boldsymbol{x}), \quad (2)$$

are minimized, where $\boldsymbol{x}$ denotes a 2D point in the image.

The OF constraints are not sufficient for the motion estimation, and in order to make the problem well-posed, one has to penalize irregular behaviour. Here we focus on the TV-L^1 energy [18], where data fidelity between two frames is measured by L^1-norms of the optical flow constraints, and the global regularization term penalizes the total variation E of the estimated motion:

$$E(v) = \int \lambda_1 \|C_T(\boldsymbol{x}, v)\| + \lambda_2 \|C_D(\boldsymbol{x}, v)\| + \|\mathscr{D}v(\boldsymbol{x})\| \, \mathrm{d}\boldsymbol{x}. \quad (3)$$

With two data terms, this energy cannot be minimized as proposed in [13], unless $\lambda_1 = 0$ (D2T) or $\lambda_2 = 0$ (T2T). However, an extension to a sum of two 1-norm data terms (including the cases $\lambda_1 = 0$ or $\lambda_2 = 0$) is presented in [19]. This solution is used to substitute the original data term solution in [13], giving an algorithm that minimizes (3). The first data term produces flows that are symmetric through the interpolated frame, while the other term allows non-symmetric motion vectors. This combination should produce motion vectors where smaller details are matched using depth information, while bigger details (including lighting changes and shadows, which are not visible from depth data) should be matched using the texture frames. With the given formulation (3) we consider three distinct cases:

T2T: $\lambda_1 = 40$, $\lambda_2 = 0$,
D2T: $\lambda_1 = 0$, $\lambda_2 = 30$,
DT2T: $\lambda_1 = 5$, $\lambda_2 = 40$.

It has to be noted that DT2T can only be calculated by using the new OF introduced here, while D2T and T2T could have been calculated by using the method presented in [13]. The final estimate of the motion v is recovered following the general implementation described by [13], with the following exceptions: 65 pyramid levels are used, and 90 warps with 10 inner iterations are performed on each level; the Gaussian smoothing of input images prior to downsampling has standard deviation 0.5 for T2T and DT2T, and 0.35 for D2T; after linear upsampling of the flows, they are filtered using a 3×3 median filter. OF naturally leads to the non pixel location problem, in which a target position in T_{t-1} and T_{t+1} does not have integer coordinates. In this case bicubic interpolation is used.

2.2. Side Information Fusion

As previously outlined, our system is based on [4], and therefore also uses the Overlapped Block Motion Compensation (OBMC) SI. We propose the use of a multi-hypothesis decoder as a way of fusing the produced SIs [14]. For each SI the probability distribution of the bits of the bitplanes is calculated using the SI and its estimated residual. The distributions are then combined together using fixed weighting coefficients. Six different coefficients are used, and the resulting distributions are fed into six LDPCA decoders. For each new received chunk of syndromes the decoding is tried; if one of the decoders converges, its result is taken as final result and its combined distribution is used to reconstruct the corresponding DCT coefficients. This process can be also seen as a decoder-based rate-optimization since the chosen solution is the one requiring less bits. We employed the 2 SIs decoder (denoted as "2SI") and a 3 SIs decoder (denoted as "3SI"). For the 2SI decoder the first SI is OBMC and the second is chosen between the ones presented here. For the 3SI decoder we use OBMC, DT2T and T2T as SIs.

Table 1: QPs used with the given quantization matrices Q_i.

Sequence \ $\mathbf{Q_i}$	Q_1	Q_4	Q_7	Q_8
Dancer	33	30	27	23
Ballet	38	31	24	19
Breakdancers	40	33	26	22

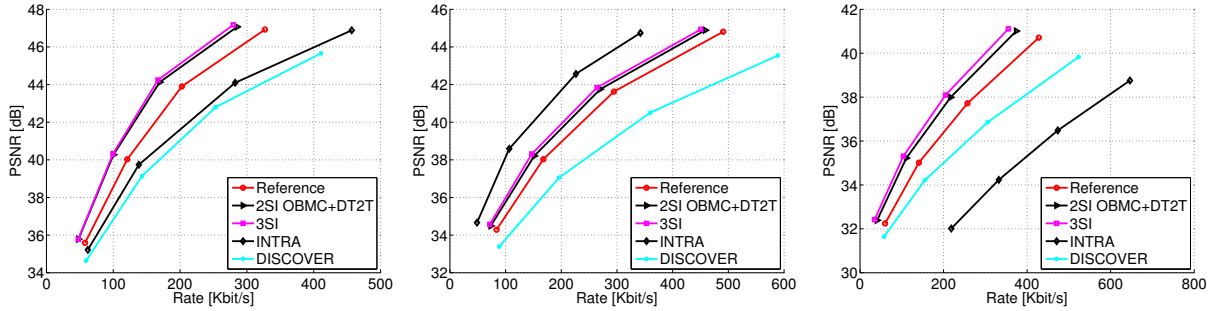
3. EXPERIMENTAL RESULTS

The system has been tested on a single view of the sequences "Breakdancers" and "Ballet" from Microsoft Research [12], and "Dancer" from Nokia Research [20]. We used the central view of the three sequences, at 15 fps downsampled to CIF resolution. The quantization matrices Q_i, $i = 1, 4, 7, 8$ of the DISCOVER project [21] are employed. The (texture) KFs are H.264/AVC Intra encoded using the QPs in Table 1. We have tested the first 100 frames of each sequence and reported the results for Group-Of-Pictures (GOP) 2, using as reference the decoder presented in [4] on texture frames. All the results and the graphs show only the WZ frames performance, since the KFs are encoded in the same manner for all the sequences. The rate of the depth frames is not taken into account, since we suppose that they are already required by the system and not only used to improve the coding performance.

The Bjøntegaard PSNR distances and bit-rate savings [22] for the 2SI decoder have been reported in Table 2, using uncompressed depth maps (denoted as $\mathrm{QP}_D = U$), and H.264/AVC Intra coded depth maps with quantization parameter $\mathrm{QP}_D = \{20, 40\}$. In Table 2, each 2SI decoder is denoted with the second employed SI, since the first one is always OBMC. DT2T is an extension of T2T, hence we report also the latter for the ease of comparison. From Table 2, in the case of uncompressed depth maps, we can see that DT2T is the best performing method for Dancer and Ballet, both medium motion sequences. For Ballet the second best method is D2T OF, while for Dancer the difference between D2T OF and T2T is negligible. It has to be noted that Ballet is a real-world sequence and depth maps have been estimated from texture data. Dancer, on the other hand, is a computer-

Table 2: Bjøntegaard Distances between the reference decoder [4] and the proposed decoders.

Sequence		T2T	$QP_D = U$			$QP_D = 20$			$QP_D = 40$		
			D2T BB	D2T OF	DT2T	D2T BB	D2T OF	DT2T	D2T BB	D2T OF	DT2T
Dancer	ΔRate[%]	15.44	11.12	17.57	23.82	9.48	17.16	24.55	5.41	12.57	19.29
	ΔPSNR[dB]	0.74	0.48	0.78	1.12	0.42	0.76	1.15	0.23	0.56	0.88
Ballet	ΔRate[%]	2.76	11.03	17.42	19.11	9.46	17.69	18.97	6.20	15.36	17.03
	ΔPSNR[dB]	0.19	0.74	1.19	1.32	0.63	1.22	1.32	0.41	1.05	1.17
Break-dancers	ΔRate[%]	3.84	8.50	12.43	11.61	7.71	12.30	11.83	5.03	11.15	10.95
	ΔPSNR[dB]	0.24	0.52	0.76	0.71	0.47	0.75	0.73	0.30	0.68	0.67



(a) Ballet (b) Breakdancers (c) Dancer

Fig. 4: RD curves, WZ frames only, uncompressed depth maps.

generated sequence in which depth maps have been generated using the actual distances of the 3D object models from the virtual camera. The depth maps of Dancer are smoother compared with those of Ballet, hence the SI of Dancer has lower quality compared with Ballet. Nevertheless, the novel DT2T approach outperforms both D2T OF and T2T, improving over the single SI decoder [4] by up to 1.32 dB. Breakdancers shows a much higher temporal activity making the motion estimation more difficult. In this case D2T OF greatly outperforms T2T. DT2T shows a negligible performance loss compared with D2T OF. In all the aforementioned cases D2T BB is not able to achieve the same performance as D2T OF due to the lack of flexibility of the block-based approach. The proposed OF-based methods show high resilience to the quantization noise of the depth maps: the performance in the case of $QP_D = 20$ are basically the same as in the uncompressed case; while in the case of $QP_D = 40$ we can notice a performance degradation, but the DT2T method is still able to achieve improvements ranging from 0.67 to 1.17 dB over [4]. It may also be noted that DT2T works correctly even when T2T has better performance compared with D2T OF (see Dancer, $QP_D = 40$) and it is still superior to the best of them. For what concerns the 3SI decoder (Table 3), it is able to correctly fuse the SIs leading to good and robust improvements, always superior to any 2SI decoder, with gains ranging from 0.90 dB to 1.49 dB. No case-specific optimization has been performed, i.e. the parameters used for the OF-based methods are fixed for all the sequences and for all the QP_D values. The RD-curves for the three sequences in the case of uncompressed texture frames are also depicted in Fig. 4,

Table 3: Bjøntegaard Distances between the reference decoder [4] and the 3SI decoder.

Sequence \ QP_D		U	20	40
Dancer	ΔRate[%]	30.69	30.80	28.40
	ΔPSNR[dB]	1.48	1.49	1.35
Ballet	ΔRate[%]	20.70	20.63	18.96
	ΔPSNR[dB]	1.45	1.45	1.32
Break-dancers	ΔRate[%]	15.15	15.05	14.49
	ΔPSNR[dB]	0.94	0.93	0.90

where the performance of [4] is denoted as "Reference". As it can be seen the decoder in [4] is able to greatly outperform the DISCOVER [6] decoder on textures in all the settings, making it more fair to compare the proposed systems with the one in [4].

4. CONCLUSION

In this work we investigated the possibility of using depth maps for improved SI generation in single-view video-plus-depth DVC. The proposed system is able to achieve good and robust improvements over one of the best single SI DVC decoders available in literature [4], with improvements ranging from 0.90 dB to 1.49 dB. OF-based methods showed clear superiority to conceptually similar block-based methods. The DT2T method was able to successfully combine the symmetrical OF approach [16] and the D2T approach introduced in this work. Finally, the multi-hypothesis decoder was able to successfully and robustly fuse the SIs here presented.

5. REFERENCES

[1] S. Mehrotra, Z. Zhang, Q. Cai, C. Zhang, and P.A. Chou, "Low-complexity, near-lossless coding of depth maps from Kinect-like depth cameras," in *Proc. of IEEE MMSP*, October 2011, pp. 1 –6.

[2] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. 19, no. 4, pp. 471 – 480, July 1973.

[3] A.D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inform. Theory*, vol. 22, pp. 1–10, 1976.

[4] X. Huang and S. Forchhammer, "Cross-band noise model refinement for transform domain Wyner-Ziv video coding," *Signal Processing: Image Communication*, vol. 27, no. 1, pp. 16 – 30, 2012.

[5] B. Girod, A.M. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," in *Proc. of the IEEE*, January 2005, vol. 93, pp. 71 –83.

[6] X. Artigas, J. Ascenso, M. Dalai, S. Klomp, D. Kubasov, and M. Ouaret, "The DISCOVER codec: Architecture, techniques and evaluation," in *Proc. of PCS*, November 2007.

[7] D. Varodayan, A. Aaron, and B. Girod, "Rate-adaptive codes for distributed source coding," *EURASIP Signal Processing Journal*, vol. 86, no. 11, pp. 3123–3130, November 2006.

[8] M. Winken, H. Schwarz, and T. Wiegand, "Motion vector inheritance for high efficiency 3D video plus depth coding," in *Proc. of PCS*, May 2012, pp. 53 –56.

[9] G. Petrazzuoli, M. Cagnazzo, F. Dufaux, and B. Pesquet-Popescu, "Wyner-Ziv coding for depth maps in multiview video-plus-depth," in *Proc. of IEEE ICIP*, September 2011, pp. 1817 –1820.

[10] M. Salmistraro, M. Zamarin, L. L. Rakêt, and S. Forchhammer, "Distributed multi-hypothesis coding of depth maps using texture motion information and optical flow," in *Proc. of IEEE ICASSP*, May 2013, *accepted*.

[11] X. Artigas, E. Angeli, and L. Torres, "Side information generation for multiview distributed video coding using a fusion approach," in *Proc. of NORSIG 2006*, June 2006, pp. 250 –253.

[12] L.C. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 600–608, 2004.

[13] L.L. Rakêt, L. Roholm, A. Bruhn, and J. Weickert, "Motion compensated frame interpolation with a symmetric optical flow constraint," in *Advances in Visual Computing*, George Bebis *et al.*, Ed., vol. 7431 of *Lecture Notes in Computer Science*, pp. 447–457. Springer Berlin Heidelberg, 2012.

[14] X. Huang, L.L. Rakêt, H.V. Luong, M. Nielsen, F. Lauze, and S. Forchhammer, "Multi-hypothesis transform domain Wyner-Ziv video coding including optical flow," in *Proc. of IEEE MMSP*, October 2011, pp. 1 –6.

[15] H.V. Luong, L.L. Raket, X. Huang, and S. Forchhammer, "Side information and noise learning for distributed video coding using optical flow and clustering," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4782 –4796, December 2012.

[16] L.L. Rakêt, J. Søgaard, M. Salmistraro, H. V. Luong, and S. Forchhammer, "Exploiting the error-correcting capabilities of low density parity check codes in distributed video coding using optical flow," in *Proc. of SPIE*, 2012, vol. 8499, pp. 84990N–84990N–15.

[17] Y. Nie and K.-K. Ma, "Adaptive rood pattern search for fast block-matching motion estimation," *IEEE Trans. Image Process.*, vol. 11, no. 12, pp. 1442 – 1449, December 2002.

[18] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime TV-L^1 optical flow," in *Ann. Symp. German Association Patt. Recogn*, 2007, pp. 214–223.

[19] A. Wedel, T. Pock, J. Braun, U. Franke, and D. Cremers, "Duality TV-L^1 flow with fundamental matrix prior," in *Image and Vision Computing*, Auckland, New Zealand, November 2008, pp. 1–6.

[20] "Extension of existing 3DV test set toward synthetic 3D video content," ISO/IEC JTC1/SC29/WG11, Doc. M19221, Daegu, Korea, January 2011.

[21] "DISCOVER project test conditions," December 2007, http://www.img.lx.it.pt/ discover/test_conditions.html.

[22] G. Bjøntegaard, "Calculation of average PSNR differences between RD-curves," in *VCEG Meeting*, Austin, USA, April 2001.